Leverage always-on voice trigger IP to reach ultra-low power consumption in voice-controlled devices

Voice-controlled devices will continue to gain new market segments in the electronics industry in the coming years because voice technologies open new horizons for the user experience. In low-power smart devices – Smartphones, Smartwatches, TV remote controls –, two functionalities are generally combined to perform voice control:

- Voice Activity Detection (VAD):
  The VAD function runs in the always-on domain. Its role, in real handsfree applications, is to wake-up a subset of the system as soon as voice is detected. As the other blocks stay in deep sleep operation during the "always listening mode", it results in a drastic reduction of the power consumption for the audio subsystem.
- Voice recognition:
  The voice recognition function, often part of an audio DSP, analyzes the audio signal and turns the recognized instructions (keywords or full speech meaning) into commands.

Due to the lack of thorough specifications to assess the performance of voice detection solutions, it is a real conundrum for users to evaluate the best solution among a jungle of true and false detection claims and without any benchmark.

The aim of this article is to help IP procurement managers, SoC architects and system makers using voice detection IP in their systems to assess and compare the detection performances.

This article also gives some clarifications on triggering solutions for voice-controlled systems and provides means to specify and compare existing solutions thanks to the MIWOK™ benchmarks.

Which characteristics to rely on to detect a voice in voice-controlled applications?

You may think, and you would be right, that the first quality of a voice activity detector is to catch any voice event while skipping all noise events. But how is it translated into reliable figures and specification?

Whereas the speech recognition community specifies recognition performance as the Word Error Rate (WER), especially for Speech To Text (STT), the voice detection community has not yet created its own metrics for voice-controlled applications.

The challenge is to create representative figures of realistic usage and situations. Developers need to consider the following parameters that influence voice detection algorithm performances:

- Input words characteristics (voice power level, language, pronunciation, accent, tone or stress characteristics),
- Environment characteristics (background power level, type of noise: wind, cars, conversation).

The voice detection functionality is different from the voice recognition functionality, as voice triggering consists only in sending an interrupt signal. This signal is a flag used to wake up parts of a system during voice activity. The most relevant performances to characterize the quality of voice detection are:

- The capability to detect the beginning of a voice event
- The capability to avoid the detection of a noise event
- The latency of detection

In low-power applications with voice capabilities, such as voice-activated remote controls, the recognition circuitry can take advantage of a real-time voice trigger (no need for data storage) to enable the best power optimization with the lowest silicon area.

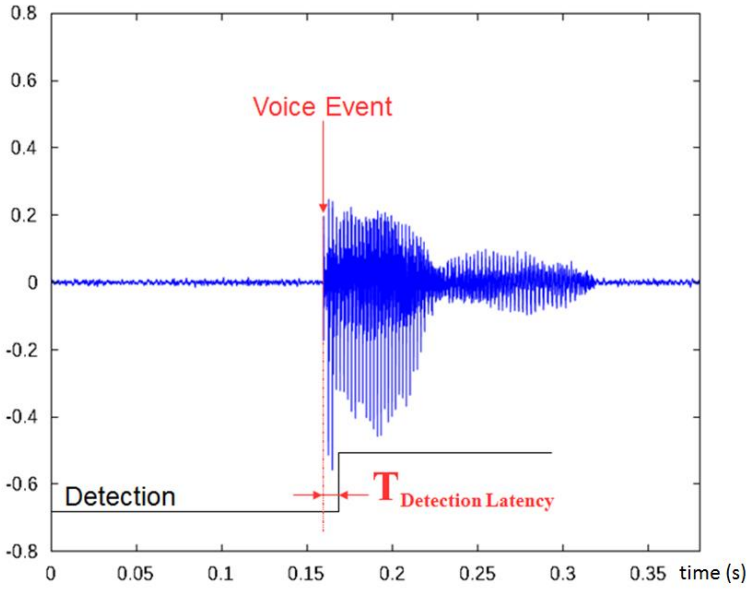

Figure 1: Latency of the voice detection real-time algorithm

Detection latency vs. language characteristics

To ensure the best recognition performances, the voice trigger must be able to detect voice activity as fast as possible. Only a minimum percentage of the first phoneme (i.e. perceptual distinct units of sound that distinguish one word from another) can be missed without degrading recognition performances.

Phonemes have diverse lengths and forms. That is the reason why keywords are often chosen for their spelling characteristics. The following figure shows how the “OK google” key-word is recognized when its first phoneme has been shortened.

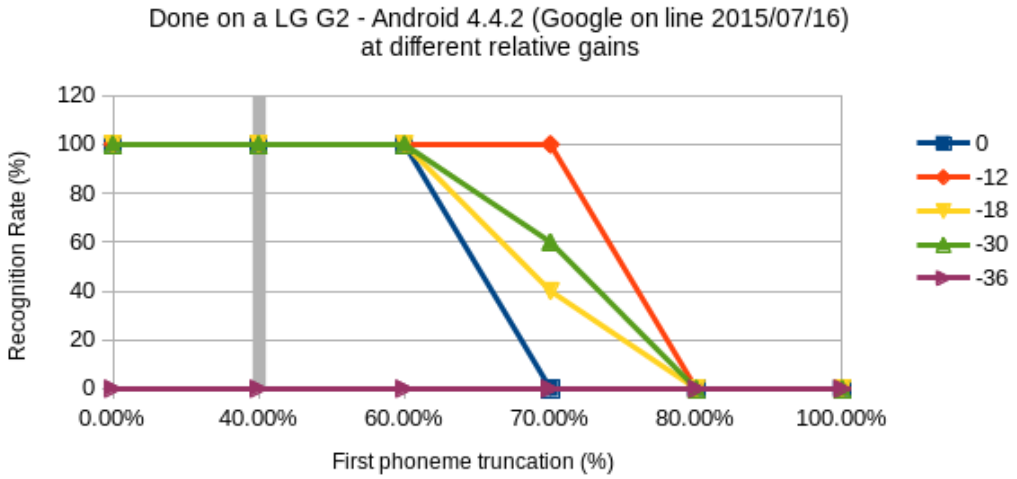

Figure 2: Percentage of key-word recognition depending on first phoneme truncation

As soon as the voice triggering system is coupled with a recognition algorithm, the first phoneme should be shortened as little as possible in order to maintain the full performance of the recognition algorithm. Therefore, a “Pass” or “Fail” criterion must be defined as the minimum detection latency that represents an acceptable amount of time before degrading the recognition performances. Regarding the impact on voice processing, the detection latency has to be expressed as a percentage of the first phoneme.

According to the tests made by Dolphin Integration’s sound experts on several recognition algorithms, the “Pass” criterion has been defined at less than 40% of the first phoneme’s duration. In this context, the first phoneme keeps its characteristics, remains intelligible by a voice-processing algorithm and the word is not degraded.

And what about unwanted detections?

Another challenge is to properly define the false detection rate. A false detection occurs due to the background sounds, without any speech. Thus, it would not be relevant to relate the number of false detections to a total number of words as this figure would be distorted proportionally to the density of speech, false detection being hidden by voice activity.

It is more relevant to state the false detection rate as the percentage of time when the application is uselessly awoken in presence of noise. On the one hand this figure gives a more relevant image of the false detections rate, on the other it also informs on the impact on power consumption at the system level.

We can summarize the main voice triggering performances as:

Voice Detected as Voice

$$VDV = \frac{Number\ of\ detected\ words*}{Total\ number\ of\ words}$$

Noise Detected as Voice

$$NDV = \frac{Duration\ of\ false\ detections**}{Total\ duration\ of a\ given\ benchmark}$$

Voice Trigger Efficiency

$$VTE = 0.5 \times VDV + 0.5 \times (1 - NDV)$$

** Word detected within 40% of its first phoneme*
*** False detection: no voice event occurs but a voice event is detected*

How to evaluate, specify and compare voice triggers

Today, no standard, voice benchmarks or reference testbenches have been adopted. Whenever a fabless company needs to acquire a voice trigger solution, the buyer cannot rely on standard figures to evaluate the performances, neither to fairly compare different solutions available on the market. The user has no other choice than to test its system in real conditions.

To deal with this issue, some methodologies already exist, which include noises and sentences, but they mainly focus on different usage or application. Output figures are given as speech hit-rate and non-speech hit-rate for all the given words inside the sentences, which is not relevant for a voice trigger application. Indeed, the performances must rely on the detection of the first word at the beginning of a speech or on specific keywords, and more precisely on the first phoneme of these words.

The first voice trigger corpus of benchmarks: MIWOK™

With more than 25 years of experience in advanced audio, Dolphin Integration drives voice triggering forward with the release of its own corpus of benchmarks – MIWOK™ –specifically developed to evaluate and compare voice triggering solutions running in always-on mode.

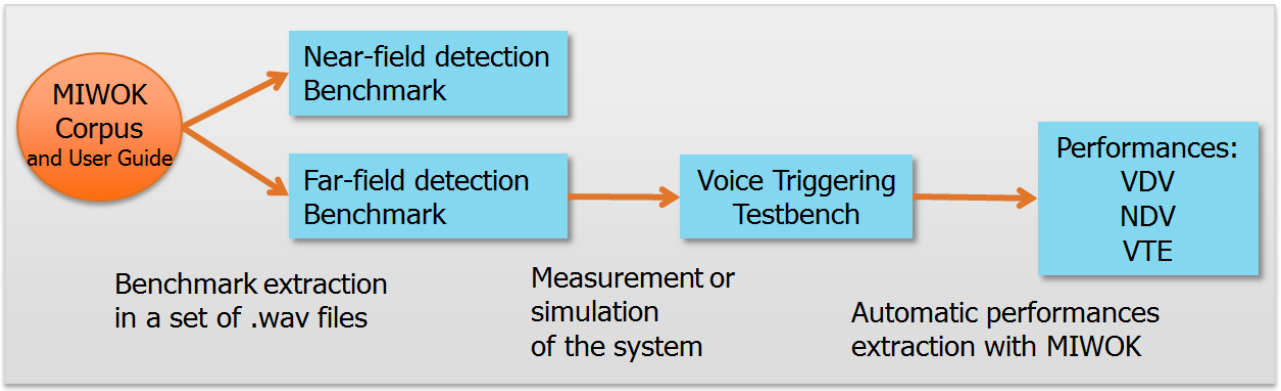

Figure 3:Test of voice triggering applications with MIWOK™

The MIWOK™ Corpus contains a set of words, background noises and tools, freely available through the Dolphin Integration's website, which aims at providing reference benchmarks and procedures in order to evaluate the performances of any voice triggering system.

Its first release was in Chinese language. It includes more than one hundred words, recorded by three different people, and more than a dozen background noises (i.e. street, airport, station, home, car…). The selection of the application context (Near-field detection or Far-field detection) and the mix of words vs. background noises is automatically computed from a GUI.

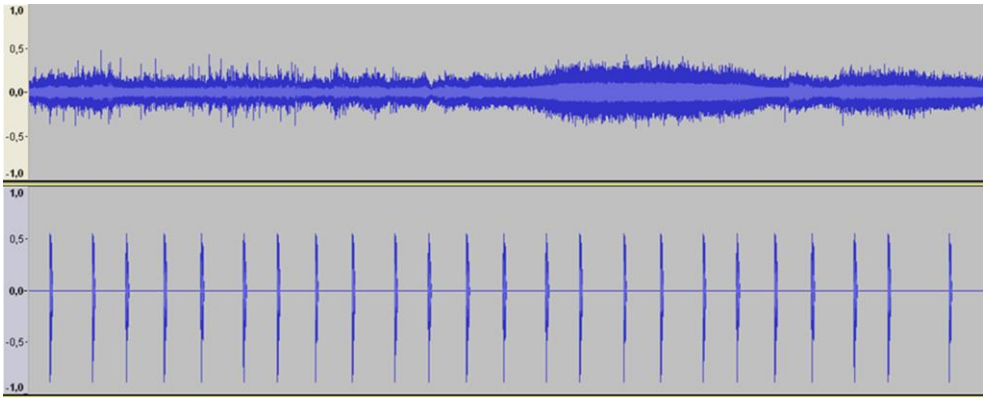

Figure 4: Benchmark input is the mix of two .wav frames: background noise (up) and word sample (down)

The MIWOK™ Corpus also provides a comprehensive methodology and detailed instructions to run simulations or measurements in a reproducible and fair manner. It also enables to easily extract the performance figures. Statistic data regarding true detections and false detections are sorted for each .wav pattern and final figures are given for the whole benchmark.

Dolphin Integration offers high performance voice trigger IPs

Dolphin Integration is pushing the boundaries of voice triggering by offering two types of voice trigger IP as hard IP. This type of implementation represents a breakthrough on the market in opposition to standard DSP-based voice detection features. The combination of a positioning early in the acquisition chain with real time detection results in an extra-low power consumption.

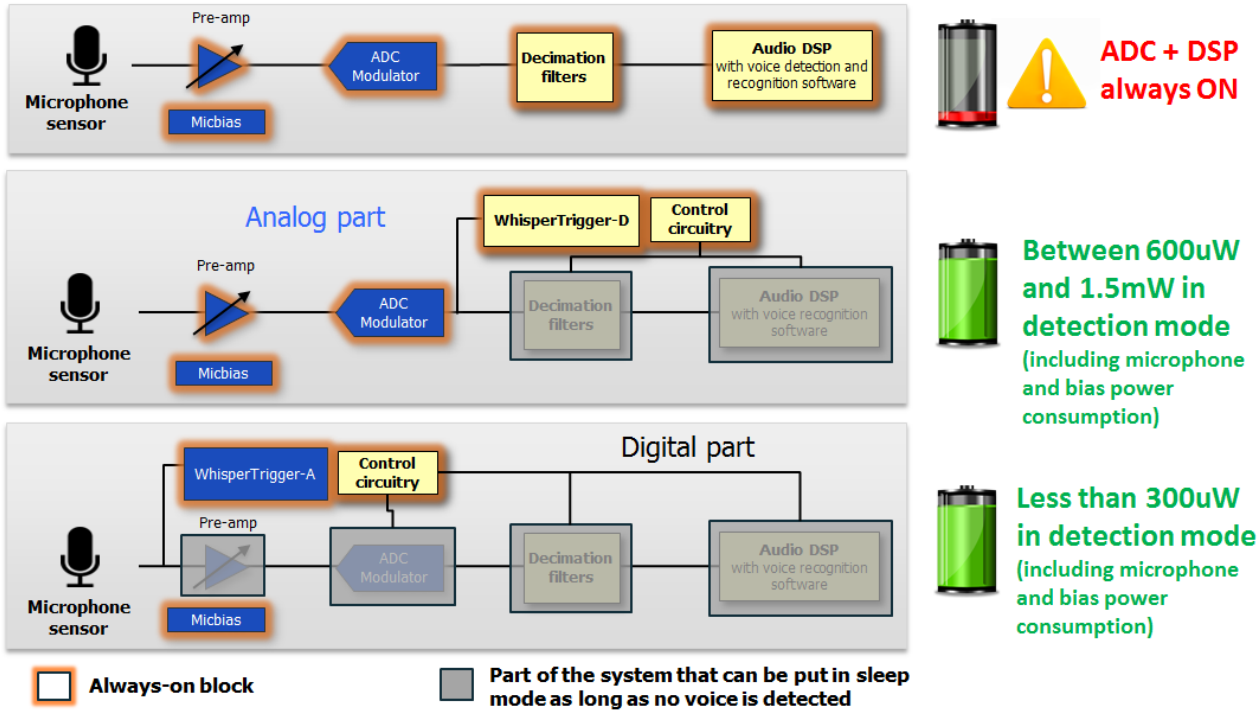

Figure 5: Architecture comparison of voice triggering solutions.

1) Digital voice trigger: WhisperTrigger™-D

This voice trigger is suitable to interface with a digital PDM microphone or to any PDM output of an audio converter. It is a unique algorithm that takes benefit of the best "state of the art" voice activity detection techniques.

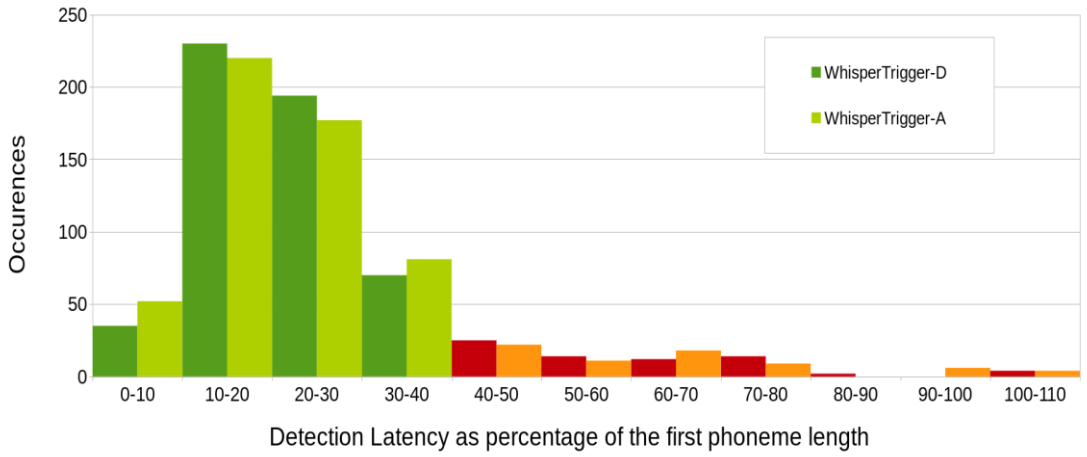

Figure 6: Distribution of the detection latency as percentage of the first phoneme (Far-field detection benchmark of MIWOK-C).

2) Analog voice trigger: WhisperTrigger™-A

This voice trigger is suitable to interface with an analog microphone (ECM or MEMS). During

always-on operation, the WhisperTrigger™-A enables to save the power consumption of the whole audio recording chain. In other words, the mixed-signal audio chain (amplifier and ADC) and the DSP can stay in sleep mode while the voice trigger is listening.

Final results are given here for the "Far-field detection" (e.g. STB, DTV) benchmark of MIWOK-C:

	WhisperTrigger-A	WhisperTrigger-D
VDV	85.64%	85.77%
NDV	2.02%	7.03%
VTE	91.81%	89.37%

Conclusion

Meeting today's low power challenges for applications embedding a voice-awakening feature requires specific performances and well-defined test parameters. The final specification is a weighted figure between detection latency, false detection rate and true detection rate.
To help IC designers and IP procurement managers assess and compare voice trigger solutions, Dolphin Integration commits to offering the first corpus of benchmarks – MIWOK™ – dedicated to this functionality.

Dolphin Integration's voice triggers have been tested with the proposed benchmarks and have been successfully embedded in several ICs of the market for a broad range of applications.

Explore Dolphin Integration IP here:

- sCODa96-H1-LB-IO-VD.01 (28 nm)
- sCODS100-LB-IO-N.12 (40 nm)
- sCODp-MT1-VD.02 (130 nm)

Please download for free the MIWOK-C Corpus at the following address:
http://www.dolphin.fr/index.php/silicon_ip/ip_products/triggers/overview

About the authors:

Paul Giletti - Audio product line manager, Dolphin Integration

After a brief experience in I/O and physical design (2003-2004), Paul Giletti became an analog design engineer in Dolphin Integration. Specialized in delta-sigma converters and audio power amplifiers, its R&D works mainly focus on high density and low power design for high performance audio applications.
He graduated in electronic and signal processing from the Polytechnical National Institute of Toulouse (INPT-ENSEEIHT, 2003)

Vincent Richard – Audio product marketing manager, Dolphin Integration

Vincent Richard is Product Marketing Manager for Audio IP. He joined Dolphin Integration in 2012 after a master degree in Management of Technologies and Innovation at Grenoble Business School. He is in charge of the product definition and promotion of audio and measurement silicon IPs. He holds a Microelectronic Engineering Master degree in Integrated Circuit design.

Main contributors of the MIWOK Corpus:
Laurent Sauzéat – Digital design engineer, Dolphin Integration
Julien Gilleron – Analog design engineer, Dolphin Integration